INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

End-to-End Speech Translation with Knowledge Distillation

Yuchen Liu[1,2]*, Hao Xiong*[4]*, Jiajun Zhang*[1,2]*, Zhongjun He*[4]*, Hua Wu*[4]*, Haifeng Wang*[4] *and Chengqing Zong*[1,2,3]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[4]Baidu Inc. No. 10, Shangdi 10th Street, Beijing, China

{yuchen.liu, jjzhang, cqzong}@nlpr.ia.ac.cn,
{xionghao05, hezhongjun, wu_hua, wanghaifeng}@baidu.com

Abstract

End-to-end speech translation (ST), which directly translates from source language speech into target language text, has attracted intensive attentions in recent years. Compared to conventional pipeline systems, end-to-end ST model has potential benefits of lower latency, smaller model size and less error propagation. However, it is notoriously difficult to implement such model which combines automatic speech recognition (ASR) and machine translation (MT) together. In this paper, we propose a *knowledge distillation* approach to improve ST by transferring the knowledge from text translation. Specifically, we first train a text translation model, regarded as the teacher model, and then ST model is trained to learn the output probabilities of teacher model through knowledge distillation. Experiments on English-French Augmented LibriSpeech and English-Chinese TED corpus show that end-to-end ST is possible to implement on both similar and dissimilar language pairs. In addition, with the instruction of the teacher model, end-to-end ST model can gain significant improvements by over 3.5 BLEU points.

Index Terms: Speech recognition, Speech translation, Knowledge distillation, Transformer

1. Introduction

Conventional speech translation system is a pipeline of two main components: an automatic speech recognition (ASR) model which transcribes source language utterances, and a text machine translation (MT) model which translates the transcripts into target language [1, 2, 3, 4, 5, 6]. This pipeline system usually suffers from time delay, parameter redundancy and error accumulation. In contrast, end-to-end speech translation ST is more compact and efficient. It can jointly optimize ASR and MT in one model and directly generate translations from source language utterances. Therefore, this model has become a new trend in speech translation fields [7, 8, 9, 10, 11, 12].

However, despite appealing advantages of end-to-end ST model, its performance is generally inferior. One reason is due to scarce data of audios paired with target translations. Previous studies resort pretraining or multi-task learning to improve the translation quality. They either pretrain ASR task on high-resource data [12], or use multi-task learning to train ST model with ASR or MT model simultaneously [10, 11]. Nevertheless, they only gain limited improvements and do not take full advantage of text data. We notice that the performance between speech translation and text translation exists a huge gap, thus how to utilize MT model to instruct ST model is of great significance.

It is a challenge to train an end-to-end ST model directly from uttrances without transcriptions while achieving comparable performance as text translation model. Considering that text translation models are prominently superior to ST model, we improve ST model by leveraging *knowledge distillation*. In knowledge distillation, there is a big superior teacher model with a small inferior student model. The student model is trained to imitate the behaviour of teacher model, such as output probabilities [13, 14], hidden representations [15, 16], or generated sequences [17], which can alleviate the performance gap between itself and the teacher model [13].

In this paper, we improve end-to-end ST model through knowledge distillation by learning knowledge from text translation model. We first train a text MT model (regarded as teacher) on parallel text data and then an end-to-end ST model (regarded as student) is trained to learn from both correction translations and the output probabilities of teacher model. Experiments on English-French Augmented LibriSpeech and English-Chinese TED corpus show that it is possible to train a compact end-to-end speech translation model on both similar and dissimilar language pairs. Furthermore, with the instruction of teacher model, end-to-end ST model can achieve significant improvements, approaching to the traditional pipeline system.

2. Related Work

End-to-end model has already become a dominant paradigm in machine translation, which adopts an encoder-decoder architecture and generates target tokens from left to right at each step [2, 4, 6, 18]. This model has also achieved promising results in ASR fields [3, 5, 19]. Recent studies purpose a further attempt to combine these two tasks together by building an end-to-end speech translation without the use of trancriptions during learning or decoding.

Anastasopoulos et al. [7] used k-means clustering to cluster repeated audio patterns and automatically align spoken words with their translations. Duong et al. [8] focused on the alignment between speech and translated phrase but not to directly predict the final translations. Bérard et al. [9] gave the first proof that end-to-end speech translation can be implemented without using any source trancriptions. They further conducted experimetns on a larger English-French dataset and proved that pretraining on ASR task can improve the performance of ST model [11]. Weiss et al. [10] used multi-task learning and

This work is done while Yuchen Liu is doing research intern at Baidu Inc.

http://dx.doi.org/10.21437/Interspeech.2019-2582

first showed that end-to-end model can outperform a cascade of independently trained pipeline system on Fisher Callhome Spanish-English speech translation task. Bansal et al. [12] found that pretraining encoder on higher-resource ASR training data can achieve significant improvements on low-resource speech translation, even when the audios in two tasks do not belong to the same language. However, these work mainly resort to pretraining acoustic encoder and do not take full advantage of text data.

Knowledge distillation is first adopted to apply for model compression, whose main idea is to train a student model to mimic the behaviors of a teacher model. It has soon been applied to a variety of tasks, like image classification [13, 20, 21, 22], speech recognition [13] and natural language processing [14, 17, 23]. The teacher and student model in conventional knowledge distillation usually handle the same task. However, in our method the teacher model and student model have different input modalities where the former input is text and the latter is audio.

3. Models

We apply end-to-end models with almost the same architecture for all three tasks (ASR, ST and MT). The architecture is adapted from *Transformer* model, which is the state-of-art model in MT task [6]. Recently, this model also begins to be used in ASR task, showing a decent performance [24, 25]. In this section, we first describe the core architecture of *Transformer* and then show how this model is applied to ASR/ST and MT task.

3.1. Core Module of Transformer

Transformer model adopts an encoder-decoder architecture with entire self-attention mechanism including scaled dot-product attention and multi-head attention. It consists of N stacked encoder and decoder layers. Each encoder layer has two blocks, which is a self-attention block followed by a feed-forward block. Decoder layer has the same architecture with encoder layer except an extra encoder-decoder attention block to perform attention over the output of the top encoder layer. Residual connection and layer normalization are employed around each block. In addition, the self-attention block in the decoder is modified with mask to prevent present positions attending to future positions during training.

Multi-head attention is applied in self-attention and encoder-decoder attention blocks to obtain information from different representation subspaces at different positions. Each head is corresponding to a scaled dot-product attention, which operates on query Q, key K and value V:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \qquad (1)$$

where d_k is the dimension of the key. Then the output values are concatenated,

$$\begin{aligned}\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \text{Concat}(\text{head}_1,\cdots,\text{head}_\text{h})\mathbf{W^O} \\ \text{where head}_\text{i} &= \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V)\end{aligned} \qquad (2)$$

where the $\mathbf{W}_i^Q \in \mathbb{R}^{d_{model}\times d_q}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_{model}\times d_v}$ and $\mathbf{W}^O \in \mathbb{R}^{d_v\times d_{model}}$ are projection matrices. $d_q = d_k = d_v = d_{model/h}$, h is the number of heads.

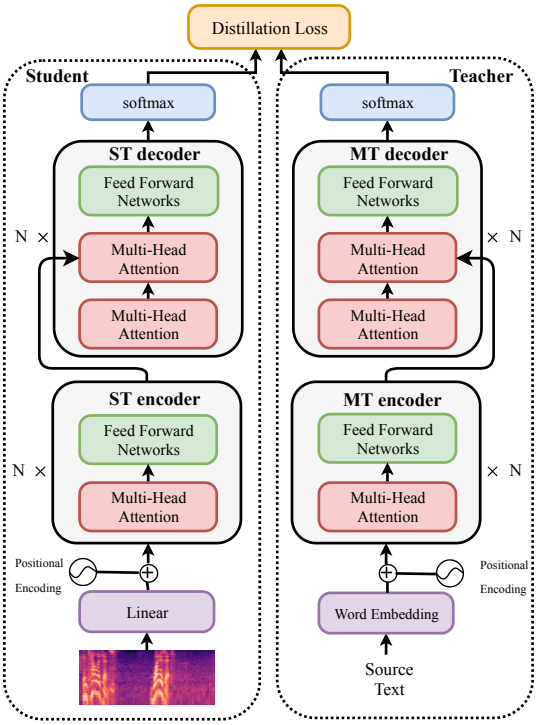

Figure 1: *Model architecture of our method. The left part is a ST model, regarded as a student model, whose input is speech. The right part is MT model, regarded as a teacher model, whose input is the source transcription. The top part is distillation loss, where the student model learns from not only the correct texts, but also the output probabilities of the teacher model.*

Position-wise feed-forward block is composed of two linear transformations with a ReLU activation in between.

$$\text{FFN}(x) = max(0, x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \qquad (3)$$

where the weights $\mathbf{W}_1 \in \mathbb{R}^{d_{model}\times d_{ff}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{ff}\times d_{model}}$ and the biases $\mathbf{b}_1 \in \mathbb{R}_{d_{ff}}$, $\mathbf{b}_2 \in \mathbb{R}^{d_{model}}$.

For the sake of brevity, we refer readers to [6] for additional details of the architecture.

3.2. ASR/ST Model

The ASR/ST model is shown in the left part of Figure 1, whose input is a series of discrete-time speech signal. We first use log-Mel filterbank to convert raw speech signal into a sequence of acoustic features and then apply mean and variance normalization. To prevent the GPU memory overflow and produce approximate hidden representation length against target length, we apply frame stack and downsample similar to [26, 27]. The final acoustic feature sequence is $S = (s_1, s_2, \cdots, s_n)$ with dimension of $d_{filterbank} \times num_{stack}$. Then the feature sequence is fed into a linear transformation with a normalization layer to map with model dimension d_{model}. In addition, positional encodings are added to the feature sequence in order to enable the model to attend by relative positions. This sequence is treated as the final input. Other parts remain the same as original *Transformer* model. For ASR task the target output is source language text, and target translation text for end-to-end ST task.

3.3. MT Model

We also use *Transformer* to train a baseline MT model, as shown in the right part of Figure 1. The difference between MT model and ASR/ST model is the input to the encoder. In MT model, $\mathrm{X} = (x_1, x_2, \cdots, x_n)$ is a sequence of tokens, representing source sentence. We embed the words in sequence *X* into a real continuous space with the dimension of $\mathbb{R}^d_{model}$, which can be fed into a neural network.

3.4. Knowledge Distillation

Training an end-to-end ST model is more difficult than MT model. The accuracy of the latter is usually much higher than the former. Therefore, we present MT model as a teacher to teach ST model. Here we give a description of the idea of knowledge distillation.

Denote $D = (s, x, y)$ as the corpus of triple data including audio, transcription in source language and its translation. The log-likelihood loss of ST model can be formulated as follows:

$$L_{\mathrm{ST}}(D;\theta) = -\sum_{(s,y)\in D} \log P(y|s;\theta) \tag{4}$$

$$\log P(y|s;\theta) = \sum_{t=1}^{N}\sum_{k=1}^{|V|} \mathbb{1}(y_t = k)\log P(y_t = k|y_{<t}, s;\theta) \tag{5}$$

where s is the acoustic feature sequence of source speech signal, y is the target translated sentence, N is the length of the output sequence, $|V|$ is the vocabulary size of the output language, y_t is the t-th output token, $\mathbb{1}(y_t = k)$ is an indicator function which indicates whether the output token is equal to the ground-truth.

We denote the output distribution of teacher model for token y_t as $Q(y_t|y_{<t}, x;\theta_T)$, and x is the source transcribed sentence which corresponds to speech signal s. Then the cross entropy between the distributions of teacher and student is:

$$L_{\mathrm{KD}}(D;\theta,\theta_T) = -\sum_{(x,y)\in D}\sum_{t=1}^{N}\sum_{k=1}^{|V|} Q(y_t = k|y_{<t}, x;\theta_T) \log P(y_t = k|y_{<t}, x;\theta) \tag{6}$$

During distillation, the student model not only learns from correct texts, but also the output probabilities of teacher model, which is more smooth and yields smaller variance in gradients [13]. Then the total loss function is,

$$L_{\mathrm{ALL}}(D;\theta;\theta_T) = (1-\lambda)L_{\mathrm{ST}}(D;\theta) + \lambda L_{\mathrm{KD}}(D;\theta,\theta_T) \tag{7}$$

where λ is a hyper-parameter to trade off two loss terms.

4. Experiments

4.1. Datasets

We conduct experiments on Augmented LibriSpeech which is collected by [28] and available for free. This corpus is built by automatically aligning e-books in French with English translations in LibriSpeech [29], which contains 236 hours of audio in total. They provide quadruplets: English audios, English transcriptions, French text translations from alignment of e-books and *Google Translate* references. Following [11], We only use the 100 hours clean train set for training, with 2 hours development set and 4 hours test set, which corresponds to 47,271, 1,071 and 2,048 sentences respectively. To be consistent with their settings, we also double the training size by concatenating the aligned references with the Google Translate references.

To verify whether end-to-end ST model can handle dissimilar language pairs, we build a corpus in English-Chinese direction. Raw data (including video, subtitles and timestamps) are crawled from TED website[1]. Audio files in each talk are extracted from video by ffmpeg[2] and saved in *wav* format. We divide each audio file into small segments based on timestamps instead of voice activity detection (VAD), because it eliminates the influence of improper fragments and guarantees each utterance containing complete semantic information, which is important for translation. In the end, we totally obtain 317,088 utterances (∼542 hours). Development and test sets are split according to the partition in IWSLT. We use dev2010 as development set and tst2015 as test set, which has 835 utterances (∼1.48 hours) and 1,223 utterances (∼2.37 hours) respectively. The remaining data are put into training set. This dataset is available on `http://www.nlpr.ia.ac.cn/cip/dataset.htm`.

4.2. Experimental Setup

Our acoustic features are 80-dimensional log-Mel filterbanks extracted with a step size of 10ms and window size of 25ms, extended with mean subtraction and variance normalization. The features are stacked with 3 frames to the left and downsample to a 30ms frame rate. We lowercase all the texts, tokenize and apply normalize punctuations by Moses [3]. For Augmented LibriSpeech corpus, we apply BPE [30] on the combination of English and French text to obtain subword units. The number of merge operations in BPE is set to 8K, resulting in a shared vocabulary with 8,159 subwords. For TED English-Chinese, the merge number is 30K, and the vocabulary size is 28,912 and 30,000, respectively. We report case-insensitive BLEU scores [31] by *multi-bleu.pl* for the evaluation of ST and MT tasks and word error rates (WER) for ASR task.

Because the size of Augmented LibriSpeech is relatively small, we set the hidden size $d_{model} = 256$, the filter size in feed-forward layer $d_{ff} = 1024$, the head number $h = 8$, the residual dropout and attention dropout are 0.1. For TED English-Chinese, we set the hidden size $d_{model} = 512$ with the filter size $d_{ff} = 2048$. MT model, as a teacher model, can use bigger parameters. We use 512 hidden sizes, 2048 filter sizes with 8 heads.The number of encoder layers and decoder layers in above models are all set to 6. We train our models with Adam optimizer [32] on 2 NVIDIA V100 GPUs.

4.3. Results

Table 1 shows the results of ASR and MT tasks on Augmented LibriSpeech. It can be seen that *Transformer* model has superior performances on both tasks, with 0.92 WER reduction and 4.1 BLEU scores improvement compared to [11]. We contribute it to the superior performance of *Transformer* model which has the ability to model long distance in sequence-to-sequence tasks, especially for MT tasks. Contrary to [11] who uses characters as output units, we consider subword units can also obtain improvements.

For ST task, we have four settings. The first is a pipeline system which uses ASR outputs as the inputs to MT model. The

[1] `https://www.ted.com`
[2] `http://ffmpeg.org`
[3] `https://www.statmt.org/moses/`

Table 1: *ASR and MT results on test set of Augmented LibriSpeech.*

LibriSpeech	Method	WER(↓)	BLEU(↑)
Bérard [11]	greedy	19.9	19.2
	beam search	17.9	18.8
Ours	greedy	21.46	21.35
	beam search	**16.98**	**22.91**

Table 2: *ST results on Augmented LibriSpeech. KD denotes knowledge distillation.*

LibriSpeech	Method	greedy	beam	ensemble
Bérard [11]	Pipeline	14.6	14.6	15.8
	End-to-end	12.3	12.9	15.5
	Pre-trained	12.6	13.3	
Ours	Pipeline	15.75	17.85	18.4
	End-to-end	10.19	13.15	17.8
	Pre-trained	13.89	14.30	
	KD	**14.96**	**17.02**	

end-to-end model is trained on source audios paired with target translation texts. The *pre-trained* model is initialized by ASR and MT models. Knowledge distillation (KD) is our method which uses MT model as a teacher to instruct *end-to-end* ST model.

As shown in Table 2, all four settings surpass the results in [11]. Noticing that there exists a large gap between the performance of end-to-end ST model and MT model, we apply *knowledge distillation* to instruct ST model by MT model. The result shows that this method can bring significant improvements on the BLEU score which increases from 14.30 to 17.02. With the instruction of MT model, the performance gap is alleviated, approaching to the pipeline system, which demonstrates the effectiveness of our method.

We also conduct experiments on an English-Chinese dataset. Table 3 presents the results. Pipeline model combines both the ASR (WER is 15.2%) and MT models. We cannot train an end-to-end ST model from random initialization parameters, since the reordering between dissimilar language pairs is too difficult to align with frame based speech representations. Here we pretrain the encoder of ST model on ASR tasks. The BLEU of end-to-end model is 16.80, which indicates the potential to implement a compact model even on dissimilar language pairs. With knowledge distillation, it also can obtain significant improvements, proving the generality of our method.

Weiss et al. showed that end-to-end ST models outperform the baseline cascade [10]. However, in our expeiments ST models are still inferior than the cascade system. We contribute it to two reasons. First, they conduct experiments on Fisher and Callhome Speech Translation Corpus which is in similar languages [33]. Besides, ASR WER on that dataset is relatively high which heavily effects the MT model in cascade system.

4.4. Analysis

To evaluate the effect of teacher model, we explore different hyper-parameters λ of the distillation loss on Augmented LibriSpeech. With λ increasing, ST will pay more attention to teacher model. When λ equals 0, it is the original end-to-end model; when λ is 1, it will ignore correct text and only learn from the teacher. As Table 4 shown, the performance becomes better with the increasing of λ. End-to-end ST model obtains

Table 3: *MT and ST results on English-Chinese TED.*

TED	MT	Pipeline	End-to-end	KD
BLEU	27.08	22.28	16.80	19.55

Table 4: *The effect of teacher model weight on ST results.*

λ	0.0	0.2	0.4	0.6	0.8	1.0
BLEU	14.30	15.68	16.73	16.62	16.93	17.02

the best performance when it only learns the output distributions of teacher model.

We further analyze how the teacher model helps ST through visualizations of the encoder-decoder attention. Figure 2 shows an example. The attentions of ASR (a) and MT (c) models are more concentrated than ST model. The attention in ST (b) model tends to be smoothed out across input frames. However, with the help of MT model, the attention of ST model with KD (d) becomes more concentrated. For example, the speech frames $l = 45 \sim 55$ are corresponding to "was talking" in ASR (a), which can be translated to "se parlait" in French (c). The attention in ST model with KD has more weights on frames $l = 45 \sim 55$ than that in original ST model.

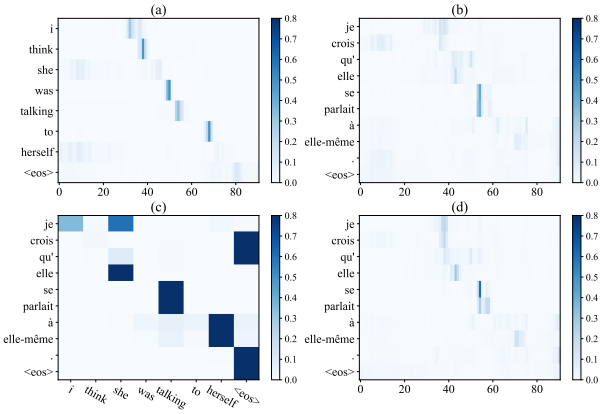

Figure 2: *The visualizations of attention in different models. (a), (b), (c), (d) are the encoder-decoder attention of ASR, end-to-end ST, MT and end-to-end ST with KD, respectively.*

5. Conclusions

In this work, we present *knowledge distillation* method to improve end-to-end ST model by transferring the knowledge from MT model. Experiments on two language pairs demonstrate that with the instruction of MT model, end-to-end ST model can gain significant improvements. Although end-to-end ST model does not outperform pipeline system, it shows the potential to come close in performance. In the future, we will utilize other knowledges like the outputs from ASR model or better MT model to further improve the performance of ST model.

6. Acknowledgements

We thank anonymous reviewers for helpful feedbacks. The research work descried in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2017YFC0822505 and the Natural Science Foundation of China under Grant No. U1836221 and 61673380.

7. References

[1] C. Zong, T. Huang, and B. Xu, "The technical analysis on automatic spoken language translation systems," *In Journal of Chinese Information Processing*, 1999.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *In Proceedings of ICLR*, 2015.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *In Proceedings of ICASSP*, 2016.

[4] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, and e. a. Mohammad Norouzi, *Googles neural machine translation system: bridging the gap between human and machine translation*. arXiv preprint arXix:1609.08144, 2016.

[5] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *In Proceedings of ICASSP*, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *In Proceedings of NIPS*, pp. 5998–6008, 2017.

[7] A. Anastasopoulos, D. Chiang, and L. Duong, "An unsupervised probability model for speech-to-translation alignment of low-resource languages," *In Proceedings of EMNLP*, 2016.

[8] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," *In Proceedings of NAACL*, 2016.

[9] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *In NeurIPS Workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[10] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *In Proceedings of Interspeech*, 2017.

[11] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," *In Proceedings of ICASSP*, 2018.

[12] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] M. Freitag, Y. Al-Onaizan, and B. Sankaran, "Ensemble distillation for neural machine translation," *arXiv preprint arXiv:1702.01802*, 2017.

[15] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," *In Proceedings of CVPR*, pp. 4133–4141, 2017.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *In Proceedings of ICLR*, 2015.

[17] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," *In Proceedings of EMNLP*, 2016.

[18] J. Zhang and C. Zong, "Deep neural networks in machine translation: An overview," *In IEEE Intelligent Systems*, pp. 16–25, 2015.

[19] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *In Proceedings of ICASSP*, 2016.

[20] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," *In Proceedings of ICCV*, pp. 1910–1918, 2017.

[21] C. Yang, L. Xie, S. Qiao, and A. Yuille, "Knowledge distillation in generations: More tolerant teachers educate better students," *In Proceedings of CVPR*, 2018.

[22] R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," *arXiv preprint arXiv:1804.03235*, 2018.

[23] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T.-Y. Liu, "Multilingual neural machine translation with knowledge distillation," *In Proceedings of ICLR*, 2019.

[24] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," *In Proceedings of ICASSP*, pp. 5884–5888, 2018.

[25] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," *In Proceedings of Interspeech*, 2018.

[26] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *In Proceedings of ICASSP*, 2015.

[27] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," *In Proceedings of ICASSP*, pp. 1–5828, 2018.

[28] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation," *In Language Resources and Evaluation*, 2018.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *In Proceedings of ICASSP*, 2015.

[30] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *In Proceedings of ACL*, 2016.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *In Proceedings of ACL*, pp. 311–318, 2002.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *In Proceedings of ICLR*, 2015.

[33] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the fisher and callhome in proceedings ofh–english speech translation corpus," *In Proceedings of IWSLT*, 2013.